社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133（10）
2003／9／14

Information Visualization System for Web Interaction

Yasufumi TAKAMA †

† Tokyo Metropolitan Institute of Technology 6-6 Asahigaoka, Hino, Tokyo 191-0065 Japan
E-mail: † ytakama@cc.tmit.ac.jp

Abstract Both Browsing and Retrieval with search engines are major operations that establish the interactions between users and the Web. Although both operations are usually combined to locate information from the Web, recent growth of the Web has overtaken the potential of this conventional interaction. This paper proposes the concept of Retrieve, Browse, and Analyze (RBA)-based interaction, as the improvement of the conventional Retrieve and Browse (RB)-based interaction. The prototype interface based on RBA-based interaction is also presented.

Keyword Information Visualization, Web Intelligence, Immune Network

Webインタラクションのための情報可視化システム

高間 康史†

†東京都立科学技術大学 〒191-0065 東京都日野市旭が丘6-6
E-mail: † ytakama@cc.tmit.ac.jp

あらまし ブラウジングと検索エンジンを用いた検索は，ユーザがWebを利用する際の主要なインタラクション方法である．Webから情報を探し，収集する際，両インタラクションを組み合わせて用いるが，拡大の一途を続ける現在のWebを最大限に活用するためには，これらのみでは不十分となりつつある．本稿では従来からの検索（Retrieve），ブラウジング(Browsing)によるRBベースインタラクションの拡張として，これに分析（Analyze）フェーズを追加したRBAベースインタラクションのコンセプトを提案し，これに基づくインタフェースのプロトタイプシステムについて示す．

キーワード 情報可視化，Webインテリジェンス，免疫ネットワーク

1. Introduction

A Web information visualization system based on RBA (Retrieve, Browse, and Analyze)-based interaction is presented for assisting user's Web interaction. A Web interaction is defined as users' activities for viewing and collecting web pages with using search engines and Web browsers. There exists vast amount of information in the Web, from which a user usually gathers information without definite information needs. Therefore, it is difficult for a user to organize and understand what he or she has gathered from the Web. In this paper, we propose the concept of RBA-based interaction. The Web information visualization system proposed in this paper employs both keyword map visualization and document clustering, which present users the topic distribution over gathered document set and document clusters, respectively. Employing the immune network-based clustering algorithm, which has been already proposed, makes it possible to find relationship between document space and keyword space.

2. Related Work

Browsing and Retrieval are the major operations that users perform on the Web. Browsing is typical operations in hyperspace (i.e., the Web). In Web hyperspace, documents are linked to others by hyperlinks, and a user can move from current document to others by clicking a hyperlink. On the other hand, a user can also get a set of documents related with his/her information needs from search engines. This operation is called retrieval hereinafter.

It seems that systems that support a user's browsing operations (browsing support system) have been major approaches in early stage of web intelligence research. However, recent success of commercial search engines such as Google has let us shift from browsing to retrieval.

Although retrieval operation has potential for user-to-web interaction, current search engines have limitation of presenting results as only a list of documents. That is, getting retrieved results is just a starting point of interaction, and users have to make much effort for inves-

tigating individual pages. Therefore, browsing is still important, which is started with using the retrieved document as the seed for browsing. Of course, users often hit on a new query while browsing the retrieved results.

2.1. Browsing Support System

Browsing support systems assist users in selecting a link to follow within the current page. Typical browsing support system, such as Letzia [6], syskill&Webert [1], Webwather [2] adds the information to each link in a document, based on which a user can select the link that will lead to the popular page, or the page of interest. This kind of systems has been developed in early stage of the Web, in which most of links are static ones. According to the spread of dynamic Web and commercial search engines of huge volume, another type of support systems that visualize the partial Web hyperspace [3, 4, 13] has become popular. BookMap [4] visualizes the user's personal hyperspace of bookmark and navigation history. It is based on the facts that (1) 92% of users have their own bookmark, and (2) more than 50% of page visits are page re-visits. It employs global fisheye and zooming operations, by which the system can show the detail of the part of hyperspace, while preserving the context (global structure).

Another example of browsing support system is Comparative Web Browser (CWB), which is designed to assist users compare the contents of a site with that of another site [7]. The CWB uses two browsing displays for displaying the contents of two sites simultaneously. When a user reads a page of a site on one of the displays, the corresponding page of another site is automatically displayed on another one.

2.2. Clustering-based Information Visualization System

Compared with above-mentioned browsing support systems that handle the hyperlinked structure of the Web, the systems that support the user's retrieving process handle a set of documents that contain the query terms. As most of documents in the set have no hyperlinks to others within the set, they should be organized in other structure than hyperlinked structure. In particular, when a large number of documents are retrieved, they should be divided into closely related subsets [5, 15]. Scatter/Gather [5] and Grouper [15] employ document-clustering approach. Scatter/Gather applies the clustering method interactively, i.e., when a user select one of the generated document cluster, the selected one is further divided into several document clusters.

The clustering result is usually presented as a list, as most search engines do. Visualization technique can also be utilized for improving the user's accessibility to the generated document clusters. CardVis [7] handles the retrieved results as a graph, where vertices denote pages and edges denote the hyperlinks between these pages. As retrieved documents do not always form a single graph, several sub-graphs are generated. CardVis [7] is based on the metaphor of a pack of playing cards, and each card shows a sub-graph. Cards are arranged in the 3D space, with which a user can interact by focus+context techniques.

RF-Cone [13] generates the tree structures when the documents of a certain topic are given, based on the similarity among documents and path from root document to each document, and visualizes them with 3D RF (relationship focused) cone tree representation.

The Category Map [14] employs SOM (self-organizing map), based on which documents are mapped onto 2D category map. Each region (a group of neighboring nodes with the same concept) corresponds to the document cluster of the concept. As SOM preserves the topological properties of document space, the Category Map can show users a relationship among document clusters.

3. Web Interface for RBA-based Interaction

3.1. Concept of Retrieve, Browse and Analyze (RBA)-based Interaction

One of the essential properties of our activities in the Web is that we do not always have the topics of interest while surfing on the Web. Therefore, not only submitting relevant queries, but also evaluating the relevance of web pages is difficult for us. Through the interaction with the Web, We find the topics of interest, acquire the background knowledge about the topics, based on which the relevance of pages is evaluated. Visualizing (partial) Web hyperspace as well as document clustering can improve the interaction between a user and the Web, as shown in Section 2.

Considering the commercial success of web search engines, it is rational that we assume the following steps for locating and gathering information in the Web:

[Retrieve] Obtain a set of documents by submitting tenta-

tive query to a search engine.

[Browse] Starting from individual documents in the retrieved results, browse their neighboring pages (documents) and collect (save) the relevant documents.

We call the interaction based on these two steps RB-based interaction. It should be noticed that a user cannot always evaluate the relevance of pages correctly, and the evaluation criteria frequently changes while he or she interacts with the Web. In other words, the context that affects the evaluation criteria is composed of the pages that have been gathered so far. Therefore, we claim that the "analyze" step should be combined with RB-based interaction. We call the interaction based on these three steps RBA-based interaction. Although Gershon [3] has already denoted the importance of the analyze step, in which the properties within a single page is analyzed. Our focus is on analyzing the set of gathered documents.

From this viewpoint, some of information visualization systems denoted in the previous section contribute for supporting RBA-based interaction. However, they put the analyze step between retrieve step and browse step. That is, the visualized space by browsing support systems is mainly used for users to browse the hyperspace. For example, the space visualized by clustering-based information visualization systems helps user explore the retrieved space. On the other hand, we propose to visualize the set of documents that is gathered as a result of the user's RB-based interaction.

3.2. System Architecture

Document clustering-based visualization is employed as our proposed system, because it is assumed that a user usually gathers the pages of interest from various Web sites, and most documents have no direct hyperlinks to others. In particular, this assumption becomes valid in retrieve step.

In order for users to understand context information from the visualized results, presenting only document clusters is not enough, but the relationship among clusters should also be presented. The SOM-based visualization systems can satisfy this to some extent, but the obtained structure seems to be fixed, even if users can manipulate the visualized space with fisheye or fractal operation [14]. Furthermore, we think that the obtained document clusters should be presented to users as the lists, because the Web users are familiar with the document lists that are returned by most of search engines.

Therefore, we propose to visualize both of document and keyword space. Document clusters are presented to users as lists, while keyword space is visualized so that the relationship among document clusters can be reflected (Fig. 1). For visualizing the keyword space, we employed the keyword map [12], on which the keywords extracted from documents are arranged so that the pairs of keywords frequently appeared in the same documents can be arranged closely to each other.

The point is how to relate the keyword map with document space, and we propose a landmark-based approach, called plastic clustering method [10,11], which is described in the next subsection.

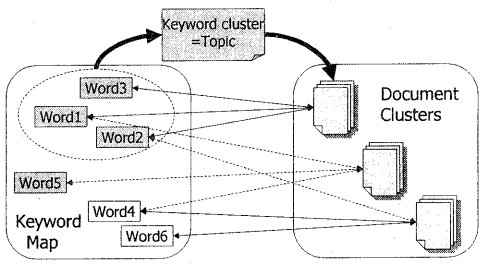

Figure 1 Correspondence between Document Space and Keyword Map

3.3. Immune Network Metaphor for Keyword Map Generation

A plastering clustering method[10, 11] has been proposed to generate a keyword map as well as document clusters. On the keyword map, the keywords related with the same topic are assumed to gather and form a cluster. The plastic clustering method extracts a representative keyword, called landmark, from each cluster. As the border of keyword clusters on the keyword map is usually not obvious, the constraints for extracting a landmark is adopted from the viewpoint of document clustering. That is, when documents containing the same landmark are classified into the same cluster, there should not exist overlapping among clusters. The algorithm of the plastic clustering method is as follows:

1. Extraction of keywords (nouns) from a document set, by using the morphological analyzer and the stop-word list. In this paper, only the keywords contained in 3 or more documents are extracted.
2. Construction of the keyword network by connecting the extracted keywords k_i to other keywords k_j or documents d_j :

   (a) Connection between k_i and k_j: (D_{ij} indicates the number of documents containing both keywords.)

   Strong connection (SC): $D_{ij} = T_k$.

   Weak connection (WC): $0 < D_{ij} < T_k$.

   (b) Connection between k_i and d_j: (TF_{ij} indicates the term frequency of k_i in d_j.)

   SC: $TF_{ij} = Td$.

   WC: $0 < TF_{ij} < T_d$.
3. Calculation of keywords' activation values on the constructed network, based on the immune network model (Eq. (1)—(5)).
4. Extraction of the keywords that activate much higher than others as landmarks after the convergence.
5. Generation of document clusters according to the landmarks

The algorithm is also shown in Fig. 2. In step4, a convergence means that the same set of keywords always becomes active (having much higher activation values (about 100 times higher in the experiments) than others [10]), which is observed after at most 1000 times calculation in most of the experiments. As for the immune network model in Step 3, the simple model that has been proposed in the field of computational biology is adopted (Eq. (1)—(5)).

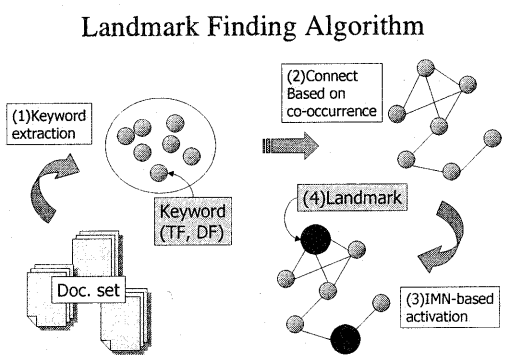

Figure 2 Landmark Finding Algorithm

$$\frac{dX_i}{dt} = s + X_i\left(f\left(h_i^b\right) - k_b\right), \tag{1}$$

$$h_i^b = \sum_j J_{ij}^b X_j + \sum_j J_{ij}^g A_j, \tag{2}$$

$$\frac{dA_i}{dt} = \left(r - k_g h_i^g\right) X_i, \tag{3}$$

$$h_i^g = \sum_j J_{ji}^g X_j, \tag{4}$$

$$f\left(h\right) = p\frac{h}{\left(h+\theta_1\right)}\frac{\theta_2}{\left(h+\theta_2\right)}, \tag{5}$$

here X_i and A_i are the concentration (activation) values of antibody i and antigen i, respectively. The s is a source term modeling a constant cell flux from the bone marrow and r is a reproduction rate of the antigen, while k_b and k_g are the decay terms of the antibody and antigen, respectively. The J^b_{ij} and J^g_{ij} ($\in\{0,\mathrm{WC},\mathrm{SC}\}$) indicate the strength of the connectivity between the antibodies i and j, and that between antibody i and antigen j, respectively. The influence on antibody i by other connected antibodies and antigens is calculated by the proliferation function (5), which has a log-bell form with the maximum proliferation rate p.

3.4. Keyword Map Visualization Interface

Keyword map-based information visualization interface is developed for visualizing the topic stream found from a sequence of document sets [12]. The developed system called TMIT (Topic Map Idea Tool) can generate keyword maps in time series. The TMIT employs the spring model [9] to arrange keywords on 2D space. Although a number of information visualization systems employ the 3D graphics, they seem to be suitable for the facilities such as museum, where visitors use the systems. We claim that the system that can be in daily use should be simple. Therefore, we employ the 2D graphics. The basic algorithm of TMIT is as follows.

1. Define the distance l_{ij} between keyword i and j based on their similarity R_{ij} by Eq. (6) (m is positive constant).

   $l_{ij} = m(1 - R_{ij})$. (6)
2. The moving distance of keyword i in each step, (δ_{xi}, δ_{yi}) is calculated by Eq. (7).

$$\left(\delta_{xi}, \delta_{yi}\right) = \left(c\frac{\partial E}{\partial x_i}, c\frac{\partial E}{\partial y_i} \right), \quad (7)$$

$$E = \sum_{i \neq j} \frac{1}{2} k_{ij} \left(d_{ij} - l_{ij} \right)^2, \quad (8)$$

$$d_{ij} = \sqrt{\left(x_i - x_j \right)^2 + \left(y_i - y_j \right)^2}. \quad (9)$$

3. In each step, the center of gravity is adjusted to the center of 2D space.

In addition to this basic algorithm, two arrangement priorities are newly introduced, i.e., the priorities based on spring constant and frictional force, respectively. It can be understood from Eq. (8) that the influence of strong spring (with large spring constant) is greater than that of weak ones. Here, the springs connecting to landmarks are given larger spring constant than others, so that the landmarks can have priority than other keywords in terms of arrangement. Furthermore, the idea of frictional force is introduced to consider the arrangement property in terms of topic stream. That is, when a new data set is to be visualized, the keyword arrangement of current keyword map should be preserved to some extent, so that users can easily grasp the relationship between the current and new maps. In TMIT, the moving distance of keyword i considering frictional force is defined as follows.

$$\left(\delta'_{xi}, \delta'_{yi}\right) = \left(\max\left(\delta_{xi} - \mu, 0\right), \max\left(\delta_{yi} - \mu, 0\right) \right). \quad (10)$$

When a new data set is to be visualized, the keyword that has already shown on the map is moved with the moving distance $(\delta'_{xi}, \delta'_{yi})$. On the other hand, the distance of newly appeared is defined by Eq. (7).

4. System Implementation

A prototype system is developed based on the description in the previous section. When designing the system, we consider the followings:

1. The system should be used by users, in combination with Web browsers for everyday use, such as IE and Netscape.
2. It should be used independent of platform (OS, hardware, etc.).
3. Further improvement or addition of new analyzing functionality should be possible in future.

Therefore, we employ server-side programming technique, as show in Fig. 3.

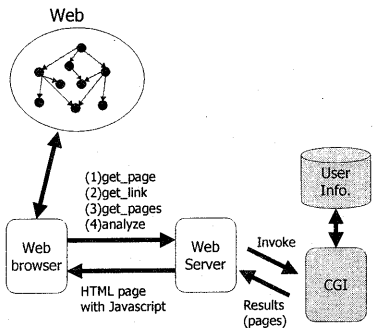

Figure 3 System Configuration

In Fig. 3, a user can interact with the Web with ordinary Web browsers as usual. The system displays a small control panel on a separate browser window, on which the user gives several instructions to CGI programs, such as follows:

[get_page] Collects the information of the page that is displayed on the user's browser window.
[get_link] Extracts and displays the link information within the displayed page.
[get_pages] The page returned by "get_link" instruction adds checkboxes to individual links, by checking which a user can collect several pages in one instruction.
[analyze] The collected page information is stored in the user information DB, to which the plastering clustering method is applied and the results including document clusters and keyword map data are returned to the user.

As the result of "analyze" instruction, the document clusters are returned as the Web page consisting of clusters with URL lists and landmarks. The page also contains the link to the data set of generated keyword map.

A user can download the data set and display it with TMIT, which is implemented with JAVA, as independent tool, not as Applet. We decide not to display the keyword map as Applet, but to provide users with the data set. The reason of providing the data set is that it consists of the

connection strength of each keyword pair, which can be utilized by users with other tools than keyword map.

Fig. 4 shows the developed system invoked from the usual Web browser. Users can access to the Web as usual, with the right-hand browser window. The control panel of the system is shown at left-hand in Fig. 4, with separate browser window. Fig. 5 shows the example of keyword map generated by the prototype system. Employing landmarks can clearly show the topic distribution.

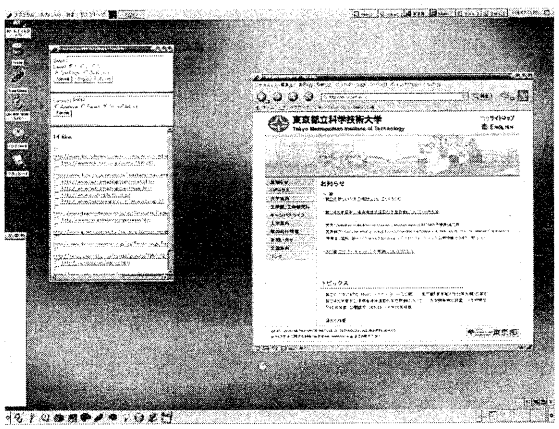

Figure 4 Developed System with Browser

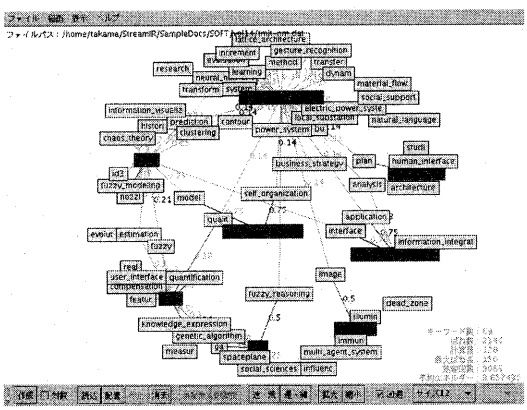

Figure 5 Keyword Map with Landmarks

5. Conclusion

The concept of Retrieve, Browse, and Analyze (RBA)-based interaction is proposed, based on which the web information visualization systems is implemented. The implemented system employs the keyword map based visualization so that users can easily understand the context of their interaction with the Web. As for the future study, the experiments with subjects are scheduled.

References

[1] M. Ackerman, et al., "Learning Probabilistic User Profiles," AI Magazine, Vol. 18, No. 2, pp. 47?56, 1997.

[2] R. Armstrong, D. Freitag, T. Joachims, T. Mitchell, "WebWatcher: A Learning Apprentice for the World Wide Web," AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, 1995.

[3] N. Gershon, J. LeVasseur, J. Winstead, J. Croall, A. Pernick, W. Ruh, "Case Study: Visualizing Internet Resources," Proc. Information Visualization (INFOVIS'95), pp. 122?128, 1995.

[4] M. Hascoet, "Interaction and Visualization Supporting Web Browsing Patterns," Proc. 5th Int'l Conf. on Information Visualization, pp.413?418, 2001.

[5] M. A. Hearst and J. O. Pedersen, "Reexaming the Cluster Hypothesis: Scatter/Gather on Retrieval Results," Proc. Of 19th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), pp. 76-84, 1996.

[6] H. Lieberman, "Letizia: An Agent That Assists Web Browsing," Proc. 14th Int'l Joint Conf. on Artificial Intelligence (IJCAI95), pp. 924?929, 1995.

[7] S. Mukherjea, Y. Hara, "Visualizing World-Wide Web Search Engine Results," Int'l Conf. on Information Visualization, p.400?405, 1999.

[8] A. Nadamoto, K. Tanaka, "A Comparative Web Browser (CWB) for Browsing and Comparing Web Pages," WWW2003, 2003.

[9] K. Takasugi, S. Kunifuji, "A Thinking Support System for Idea Inspiration Using Spring Model," J. of Japanese Society for Artificial Intelligence, Vol. 14, No. 3, pp. 495?503, 1999 (written in Japanese).

[10] Y. Takama and K. Hirota, "Consideration of Presentation Timing Problem for Chance Discovery," 5th World Multiconference on Systems, Cybernetics and Informatics (SCI2001), 8, pp. 429?432, 2001.

[11] Y. Takama and K. Hirota, "Web Information Visualization Method Employing Immune Network Model for Finding Topic Stream from Document-Set Sequence," J. of New Generation Computing, Vol. 21, No. 1, pp. 49-59, 2003.

[12] Y. Takama and Tetsuya Hori, "Application of Immune Network Metaphor to Keyword Map-based Topic Stream Visualization," Proc. 2003 IEEE Int'l Symp. on Computational Intelligence in Robotics and Automation (CIRA2003), pp. 770-775, 2003.

[13] T. Teraoka, M. Maruyama, "Research Report: Adaptive Information Visualization Based on the User's Multiple Viewpoints –Interactive 3D Visualization of the WWW—," Proc. IEEE Symposium on Information Visualization (InfoVis'97), pp. 25?28, 1997.

[14] C. C. Yang, H. Chen, K. Hong, "Internet Browsing: Visualizing Category Map by Fisheye and Fractal Views," Proc. Int'l Conf. On Information Technology: Coding and Computing (ITCC'02), pp. 34?39, 2002.

[15] O. Zamir and O. Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results," Proc. 8th International WWW Conference, 1999.